Task Set 2 — Human Semantic Evaluation of Functional Dependencies

1. Objective

The objective of Task Set 2 is to **assess the semantic validity of functional dependencies discovered algorithmically**, using **human judgment** grounded in domain knowledge and the formal definition of functional dependencies as **exception-free deterministic rules**.

Only functional dependencies provided by the algorithm were evaluated.
No functional dependencies were inferred from the data.

---

2. Evaluation Criteria

Each functional dependency was classified into one of the following categories:

- **Meaningful**: plausible as a real-world deterministic rule
- **Accidental**: holds in the dataset but unlikely to generalize
- **Encoding-based**: caused by identifiers, derived attributes, or data encoding
- **Degenerate**: right-hand side already contained in the left-hand side
- **Unlikely**: implausible as a deterministic real-world rule

Judgments focus on **determinism**, not correlation.

---

3. Dataset-Level Semantic Analysis

3.1 Iris

Selected FDs

- (Sepal length, Sepal width, Petal length) → Species
- (Sepal length, Petal length, Petal width) → Species
- (Sepal width, Petal length, Petal width) → Species

Human classification

- **Meaningful**

Explanation
Species membership in the Iris dataset is defined by combinations of morphological

measurements in a controlled botanical setting. Exact determinism is plausible and consistent with the dataset's design.

3.2 Abalone

Selected FDs

- (Viscera weight, Shell weight, Whole weight) → Shucked weight
- (Shell weight, Shucked weight, Whole weight) → Viscera weight
- (Height, Viscera weight, Shell weight, Shucked weight) → Whole weight
- (Viscera weight, Shell weight, Whole weight) → Sex
- (Shucked weight, Length, Whole weight) → Rings

Human classification

- **Encoding-based** (for weight-composition dependencies)
- **Unlikely** (for weight → Sex)
- **Accidental** (for size/weight → Rings)

Explanation
Exact dependencies between component weights and total weight strongly suggest encoding or derivation effects. Dependencies implying deterministic prediction of categorical attributes such as Sex or Rings are implausible in a biological system and are best explained as dataset-specific artifacts.

3.3 Breast-Cancer-Wisconsin

Selected FDs

- Cell size → Cell shape
- (Cell size, Marginal adhesion) → Cell shape
- (Cell size, Clump thickness) → Cell shape
- Sample ID → Cell shape
- (Sample ID, Cell shape) → Cell shape

Human classification

- **Accidental** (feature-to-feature dependencies)
- **Encoding-based** (ID-based dependencies)
- **Degenerate** (RHS included in LHS)

Explanation
While cytological features are correlated, exact functional dependence is too strong given biological variability. Dependencies involving the sample identifier reflect record identification rather than semantic rules.

3.4 Bridges

Selected FDs

- Material → Bridge type
- (Material, Span length) → Bridge type
- (Year built, Material) → Bridge type
- Bridge ID → Material
- (Bridge ID, Material) → Material

Human classification

- **Accidental** (design-related dependencies)
- **Encoding-based** (ID-based dependencies)
- **Degenerate** (RHS ⊆ LHS)

Explanation
Engineering materials and dimensions constrain design choices but do not determine them uniquely. Dependencies involving bridge identifiers are purely encoding-based and carry no semantic meaning.

3.5 Echocardiogram

Selected FDs

- Ejection fraction → Survival

- Wall motion index → Survival
- Fractional shortening → Ejection fraction
- Patient identifier → Survival
- (Survival, Age) → Survival

Human classification

- **Accidental** (clinical measurement → outcome)
- **Encoding-based** (derived measurement relationships)
- **Encoding-based** (identifier-based)
- **Degenerate** (RHS included in LHS)

Explanation
Medical measurements are predictive but not deterministically linked to outcomes. Exact dependencies involving derived cardiac metrics suggest encoding effects rather than universal clinical laws.

3.6 Hepatitis

Selected FDs

- Bilirubin → Outcome
- (Albumin, Protime) → Outcome
- (Ascites, Albumin) → Outcome
- Patient identifier → Outcome
- (Outcome, Age) → Outcome

Human classification

- **Accidental** (clinical variables → outcome)
- **Encoding-based** (identifier-based)
- **Degenerate** (RHS included in LHS)

Explanation
Clinical outcomes depend on many interacting factors and cannot be determined exactly

by a small set of measurements. Deterministic dependencies observed in the dataset are unlikely to generalize and are best interpreted as artifacts of small sample size and sparsity.

---

4. Cross-Dataset Semantic Patterns

Dataset	**Dominant Human Classification**
Iris	Meaningful
Abalone	Encoding-based / Accidental
Breast-Cancer-Wisconsin	Accidental / Encoding-based
Bridges	Accidental / Encoding-based
Echocardiogram	Accidental / Encoding-based
Hepatitis	Accidental / Encoding-based

---

5. Key Insights from Human Semantic Judgment

1. **Exact determinism is rare** in real-world biological, medical, and social systems.
2. Many algorithmically discovered FDs arise from:
   - identifiers,
   - derived attributes,
   - limited sample size,
   - or discretization.
3. Correlated attributes are often mistaken for deterministic relationships.
4. Human judgment is essential to distinguish **structural validity** from **semantic validity**.

---

6. Conclusion

Task Set 2 demonstrates that while functional dependency discovery algorithms can identify large numbers of formally valid dependencies, most of these do not correspond to meaningful real-world rules. Human semantic evaluation reveals that the majority of dependencies are accidental or encoding-based, underscoring the need for semantic filtering and hybrid approaches in dependency discovery.

Dataset	Selection type	Functional Dependency (FD)	Human Class	Justification
Iris	Plausible	(Sepal length, Sepal width, Petal length) → Species	Meaningful	Species are defined by combinations of morphological traits in a controlled botanical dataset.
Iris	Plausible	(Sepal length, Petal length, Petal width) → Species	Meaningful	Exact determinism is plausible given careful data collection and limited species.
Iris	Plausible	(Sepal width, Petal length, Petal width) → Species	Meaningful	Consistent with how species are encoded in the dataset.
Abalone	Plausible	(Viscera weight, Shell weight, Whole weight) → Shucked weight	Encoding-based	Component and total weights appear mathematically linked.
Abalone	Plausible	(Shell weight, Shucked weight, Whole weight) → Viscera weight	Encoding-based	Suggests derived or constructed measurements.
Abalone	Plausible	(Height, Viscera weight, Shell weight, Shucked	Encoding-based	Whole weight appears deterministically reconstructed from components.

Dataset	**Selection type**	**Functional Dependency (FD)**	**Human Class**	**Justification**
		weight) → Whole weight		
Abalone	Suspicious	(Viscera weight, Shell weight, Whole weight) → Sex	Unlikely	Biological sex cannot be determined deterministically from weight measures.
Abalone	Suspicious	(Shucked weight, Length, Whole weight) → Rings	Accidental	Age correlates with size but exact determinism is implausible.
Abalone	Suspicious	(Shell weight, Length) → Sex	Unlikely	Strong biological implausibility.
Breast-Cancer-Wisconsin	Plausible	Cell size → Cell shape	Accidental	Features are correlated but not deterministically linked.
Breast-Cancer-Wisconsin	Plausible	(Cell size, Marginal adhesion) → Cell shape	Accidental	Combining correlated features increases prediction, not determinism.
Breast-Cancer-Wisconsin	Plausible	(Cell size, Clump thickness) → Cell shape	Accidental	Clinical variability prevents exact functional dependence.
Breast-Cancer-Wisconsin	Suspicious	Sample ID → Cell shape	Encoding-based	Identifier uniquely references records without semantic meaning.
Breast-Cancer-Wisconsin	Suspicious	(Sample ID, Cell shape) → Cell shape	Degenerate	RHS already appears in LHS.

Dataset	**Selection type**	**Functional Dependency (FD)**	**Human Class**	**Justification**
Breast-Cancer-Wisconsin	Suspicious	(Cell size, Sample ID) → Cell size	Degenerate	Trivial dependency adding no information.
Bridges	Plausible	Material → Bridge type	Accidental	Material constrains but does not uniquely determine design.
Bridges	Plausible	(Material, Span length) → Bridge type	Accidental	Multiple bridge types remain possible under same constraints.
Bridges	Plausible	(Year built, Material) → Bridge type	Accidental	Reflects historical tendencies, not deterministic rules.
Bridges	Suspicious	Bridge ID → Material	Encoding-based	Identifier-based dependency.
Bridges	Suspicious	Bridge ID → Span length	Encoding-based	Identifier uniquely determines record attributes.
Bridges	Suspicious	(Bridge ID, Material) → Material	Degenerate	RHS included in LHS.
Echocardiogram	Plausible	Ejection fraction → Survival	Accidental	Prognostic but not deterministically linked to outcome.
Echocardiogram	Plausible	Wall motion index → Survival	Accidental	Strong predictor, not exception-free.
Echocardiogram	Plausible	Fractional shortening → Ejection fraction	Encoding-based	Likely derived or mathematically related measures.
Echocardiogram	Suspicious	Patient identifier → Survival	Encoding-based	Identifier-based dependency.

Dataset	**Selection type**	**Functional Dependency (FD)**	**Human Class**	**Justification**
Echocardiogram	Suspicious	(Survival, Age) → Survival	Degenerate	RHS already known.
Echocardiogram	Suspicious	(EF, Wall motion index) → EF	Degenerate	Trivial functional dependency.
Hepatitis	Plausible	Bilirubin → Outcome	Accidental	Clinical indicator correlates with outcome but does not determine it.
Hepatitis	Plausible	(Albumin, Protime) → Outcome	Accidental	Prognostic indicators are probabilistic.
Hepatitis	Plausible	(Ascites, Albumin) → Outcome	Accidental	Severe disease markers do not uniquely determine survival.
Hepatitis	Suspicious	Patient identifier → Outcome	Encoding-based	Identifier-based dependency.
Hepatitis	Suspicious	(Outcome, Age) → Outcome	Degenerate	RHS already contained in LHS.
Hepatitis	Suspicious	(Steroid, Antivirals) → Sex	Unlikely	No plausible deterministic relationship.